DOCUMENT RESUME

ED 227 140 TM 830 139

AUTHOR Thompson, Bruce
TITLE Two "Non-Symmetric" Methods of Canonical Analysis.
PUB DATE Nov 82
NOTE 23p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (New Orleans, LA, November 10, 1982).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; Evaluation Criteria; *Multivariate Analysis; Predictor Variables; *Research Methodology; *Statistical Analysis
IDENTIFIERS Correlation Matrices; Criterion Variables

ABSTRACT

Conventional canonical methods distinguish between the two variable sets being analyzed, but the methods do not attempt to optimize the variance from a given variable set that will be contained in the final solution. In this respect canonical methods are said the be "symmetric." This paper proposes two non-symmetric, canonical-like techniques that can be employed when theoretical or utility considerations suggest that one variable set (usually the criterion set) should be emphasized over the other variable set. The criteria that the two methods meet are discussed, as are various software considerations. (Author/CM)

TWO "NON-SYMMETRIC" METHODS OF CANONICAL ANALYSIS

Bruce Thompson
University of New Orleans 70148

Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, November 10, 1982

ABSTRACT

Conventional canonical methods distinguish between the two variable sets being analyzing, but the methods do not attempt to optimize the variance from a given variable set that will be contained in the final solution. In this respect canonical methods are said the be "symetric." The paper proposes two non-symetric, canonical-like techniques that can be employed when theoretical or utility considerations suggest that one variable set (usually the criterion set) should be emphasized over the other variable set. The criteria that the two methods meet are discussed, as are various software considerations.

Kerlinger (1973, p. 652) has suggested that "it is not easy to find research studies that have used canonical analysis. In earlier years, of course, the calculations involved were prohibitive. Today, even with computer facilities and programs available, the method is evidently not well known. This is regrettable, because some research problems almost demand canonical analysis." More recently, Thorndike (1977, p. 76) expressed similar sentiments: "Given the substantial theoretical literature on canonical analysis, it is surprising to find that the technique has seen relatively infrequent use by researchers studying substantive problems. Instances where the methods of canonical analysis have been applied rather than studied for their own sake are relatively rare (but on the increase)." Still more recently, Kettenring (1982, p. 355) noted that "since canonical analysis is usually considered to be one of the 'major methods' of multivariate analysis, it is perhaps surprising that this technique does not in fact play a larger role in data analysis."

Several possible causes for a reticience to apply canonical methods have been cited, including the complexity of canonical mathematics (Thompson, 1982, p. 467) and difficulties in interpreting canonical results (Thompson, 1980a, pp. 16-17). However, Cronbach (1971, pp. 489-490) has argued that "statistical devices such as canonical correlation, for handling handling several predictors and several criteria simultaneously, are not appropriate for the decision-oriented study... Utility depends upon values, not upon the statistical connections of scores." Kettenring (1982, p. 365) made a similar point: "To achieve its potential, better methods are needed for selecting 'canonical variables' which have practical as well as theoretical interest and for making statistical inferences about them." These views stand somewhat in contrast with Levine's (1977, p. 8) position that "especially with respect to canonical correlation, there seem to be relatively few remaining puzzles to be solved."

The purpose of this paper is to present an extension of those canonical methods which have traditionally been available to researchers. The extension will be discussed in the context of a concrete heuristic example. However, some discussion of more conventional canonical methods (Hotelling, 1935) is required to form the framework for the presentation.

The Logic of Canonical Analysis

Canonical correlation analysis is employed to study relationships between two variable sets when each variable set consists of at least two variables. Thus, Table 1 presents the data for what is the simplest canonical case, since only two criterion variables, X and Y, and only two predictor variables, A and B, are involved. Of course, these data are presented only for the purposes of discussion, since the hypothetical sample size (n = 10) is absurdly small.

INSERT TABLE 1 ABOUT HERE.

The first step in a canonical correlation analysis involves the calculation of the intervariable correlation matrix. A symetric matrix of reduced rank equal to the number of variables in the smaller of the two variable sets is then derived from the intervariable correlation matrix (see Cooley & Lohnes, 1971, p. 176 for details). This matrix is presented in Table 2. The eigenvalues of the Table 2 matrix (as some readers may wish to verify by subjecting the matrix to a principal components analysis) each represent a squared canonical correlation coefficient. Since the number of eigenvalues which can be calculated for such a matrix equals the number of rows (or columns) in the matrix, it should be clear that the maximum number of canonical correlation coefficients which can be derived for a data set equals the number of variables in the smaller of the two variable sets. Because in this case both variable sets consist of two variables, only two canonical correlation coefficients can be calculated.

INSERT TABLE 2 ABOUT HERE.

A squared canonical correlation coefficient indicates the proportion of variance which two composites derived from the two variable sets linearly share. The composites are derived by multiplying the Z scores of each person on each variable by the corresponding canonical function coefficients. Thus, a canonical correlation coefficient is simply the Pearson product-moment bivariate correlation between two linear composites derived from the two variable sets.

Table 3 provides an illustration of these calculations. The table only presents the coefficients for the first canonical function, although it has already been noted that a second function could have been calculated. The canonical correlation is the bivariate correlation between the two table columns headed "Criterion Composite" and "Predictor Composite." These 10 pairs of values are plotted in the Figure 1 scattergram. The figure also presents the regression line for the two sets of 10 composites. Since the composites are themselves also in Z score form, the regression line passes through the mean of both composites, i.e., the X-Y intercept, and the line's slope (.305) equals the bivariate correlation between the two composites and also the canonical correlation between the two variable sets.

INSERT TABLE 3 AND FIGURE 1 ABOUT HERE.

Some Non-symetric Canonical-like Methods

As the previous discussion suggests, conventional canonical methods derive equation weights which optimize the correlations between the canonical composites. Conventional methods thus give similar consideration to both variable sets when deriving function coefficients, i.e., the methods distinguish between the predictor and the criterion variable sets but do not emphasize one variable set over the other. In this respect conventional canonical methods may be said to be "symetric." These features of conventional canonical analysis are disturbing in research situations where the researcher may wish to emphasize one variable set over the other. Two "non-symetric" methods of attending to variable set distinctions will be presented here in the context of a concrete heuristic example.

The correlation matrix presented in the bottom triangle of the Table 4 matrix provides the heuristic. The data ($\underline{n}$ = 235) were reported by Thompson (1980b) and involved two sets of variables respectively of size four and 10. Table 5 presents the results obtained by analyzing the data with conventional, symetric canonical methods. The sum of the four squared canonical correlation coefficients for the Table 5 results was .593. Since conventional functions are uncorrelated, the squared correlations can be added to determine the cumulative proportion of information shared by all the possible canonical composites that could be derived from a data set in which the smallest variable set contained four variables.

<u>INSERT TABLES 4 AND 5 ABOUT HERE.</u>

Suppose, however, that the researcher wishes to derive function weights for the 10 criterion variables subject to the restriction that all of the variance in the criterion variables must be represented in the final solution. One possible way to do this is to employ criterion variable weights such as those presented in Table 6. These weights imply that *y* multiple regression analyses were conducted separately for each of the *y* criterion variables.

INSERT TABLE 6 ABOUT HERE.

The squared multiple correlations from Table 6 sum to .660. However, this result overestimates the cumulative portion of information shared by the score composites, because the Table 6 functions are not uncorrelated. The argument that these functions are correlated should seem reasonable since the criterion variable weights were designated without considering the correlations among the variables.

However, it is possible to provide a non-symetric method which (a) employs all the variance in the criterion variables (presuming, as will usually be the case, that this is the variable set of the most interest to the researcher), (b) sheds light on the structure underlying the correlations among the criterion variables, and which (c) allows uncorrelated functions to be defined, at least when the predictor variables are uncorrelated. The method requires three steps. First, the intradomain criterion correlations are subjected to a full-rank, principal components analysis, using a readily available routine from a computer package such as SPSS (TYPE=PA1). By definition, a full-rank principal components solution contains all of the variance present in the y criterion variables. Second, the orthogonal components are rotated to some orthogonal criterion (in this case Varimax), and factor scores are calculated (scores are requested in SPSS with the inclusion of "FACSCORE" on the procedure card). Finally, correlation coefficients among these factor scores and the predictor variables are computed (see upper triangle of the Table 4 matrix) and all possible y multiple regression equations are calculated. Again, routines for performing this step of the analysis are readily available. The results produced in this manner for the heuristic data are presented in Table 7. The function weights for the criterion variables are the "factor score coefficients" (obtained from SPSS during the principal components step of the analysis merely by requesting STATISTIC 7) derived by postmultiplying the inverted intradomain criterion variable correlation coefficients by the rotated principal components matrix; the function coefficients for the predictor variables are the beta weights derived from the regression analysis.

INSERT TABLE 7 ABOUT HERE.

Since in this case the four predictor variables happened to be uncorrelated, and since this non-symetric procedure always produces uncorrelated criterion composites, the functions themselves are uncorrelated. Thus, the sum of the squared multiple correlations was .593, just as the sum of the squared canonical correlation coefficients was also .593 when conventional canonical functions were computed.

A second and even more elegant non-symetric method can be proposed. This method (a) employs all the variance in the criterion variables, (b) sheds light on the structure underlying the correlations among the criterion variables, (c) allows orthogonal functions to be defined, and (d) is "confirmatory" in that _a priori_ expectations regarding the criterion variables are considered. This last element is responsive to Cronbach's (1971) previously mentioned utility concerns and can minimize the extent to which the solution capitalizes on sampling error. However, unlike the previously discussed procedures, this second method can not be implemented solely with the use of widely available statistical packages such as SPSS.

The procedure requires four steps. First, the intradomain criterion correlations are subjected to a full-rank principal components analysis. Second, the components are rotated to a position of "best fit" with an _a_ priori defined "target" matrix (for an example see Thompson & Pitts, 1981/82), typically consisting of ones, zeroes, and negative ones. This rotation can be performed using the computer program provided by Veldman (1967). Third, factor scores are computed using the least squares algorhythm:

$$Z_{N\times V} \; R^{-1}_{V\times V} \; C_{V\times F} = F_{N\times F} ,$$

where Z refers to the V scores of the N people in the scores' standardized form, R is the intervariable correlation matrix, and C is the principal components matrix derived in the previous step. The calculation of these factor scores can be facilitated at most computing facilities by the use fairly "user friendly" utility packages such as IMSL. Finally, the correlation coefficients among the y factor scores and the predictor variables are computed and all possible y multiple regression equations are computed.

Discussion

The non-symetric methods suggested here can provide, at least in some instances, both substantive and heuristic benefits when compared with conventional canonical methods. The non-symetric methods augment several very helpful extensions of conventional canonical methodology, including most notably stepwise techniques (Rim, 1972; Thompson, 1982), part and partial canonical methods (Lee, 1978), and redundancy analysis (van den Wollenberg, 1977). Of course, when the non-symetric methods are applied for the purposes of actual substantive inquiry, it becomes important to supplement the analysis by the computation of the variables' structure coefficients (see Levine, 1977, p. 20). The reader is also cautioned that experimentwise Type I error rates are inflated by the use of traditional regression test statistics with the non-symetric methods. However, when the functions are uncorrelated, as they were here, the exact experimentwise error rate would be: alpha* = 1 - k, where k equals ((1 - alpha) raised to the y power), where y is the number of criterion variables. A reasonable approach in such a case might be to test each function at the alpha/y level of statistical significance.

The non-symetric methods presented here will be most helpful when there are clear theoretical distinctions between the predictor and the criterion variable sets, and when the research situation implies that optimizing the variance of the criterion variables is at least as important as optimizing the correlation between the variable sets' composites. The non-symetric methods will also be most appropriate when there is a definite interest in exploratory (the varimax rotated solution) or "confirmatory" (the "best fit" solution) investigation of the structure underlying the the criterion variables. Thus, the techniques will be particularly potent when the criterion variables are themselves correlated (as they tended not to be in the heuristic example presented here), because then a few composites can contain the preponderance of the criterion variables' variance, and a more parsimonious solution will result.

But the non-symetric methods are also valuable to the extent that they may help to demystify conventional canonical methods. It has been noted that canonical analysis is essentially a principal components analysis of a particular matrix derived from the intervariable correlation matrix (e.g., Table 2). Similarly, the non-symetric methods relied heavily upon the use of principal components analyses. Thus it can be suggested that the symetric and the non-symetric methods are somewhat analogous. These conceptual linkages among the techniques merely suggest, as Knapp (1978) has shown, that canonical methods represent a most-general data-analytic system, and that canonical methods subsume all parametric statistical techniques.

References

Cooley, W.W., & Lohnes, P.R. Multivariate data analysis. New York: Wiley, 1971.

Cronbach, L.J. Validity. In R.L. Thorndike (Ed.), Educational measurement. Washington: American Council on Education, 1971.

Hotelling, H. The most predictable criterion. Journal of Experimental Psychology, 1935, 26, 139-142.

Kerlinger, F.N. Foundations of behavioral research (2nd ed.). New York: Holt, Rinehart & Winston, 1973.

Kettenring, J.R. Canonical analysis. In S. Kotz & N.L. Johnson (Eds.), Encyclopedia of statistical sciences (vol. 1). New York: Wiley & Sons, 1982, pp. 354-365.

Knapp, T.R. Canonical correlation analysis: a general parametric significance-testing system. Psychological Bulletin, 1978, 85, 410-416.

Lee, S. Generalizations of the partial, part and bipartial canonical correlation analysis. Psychomatrika, 1978, 43, 427-431.

Levine, M.S. Canonical analysis and factor comparison. Beverly Hills: Sage, 1977.

Rim, E. A stepwise canonical approach to the selection of "kernel" variables from two sets of variables (Doctoral dissertation, University of Illinois at Urbana-Champaign, 1972). Dissertation Abstracts International, 1973, 34, 623A. (University Microfilms No. 73-17,386)

Thompson, B. Canonical correlation: recent extensions for modelling educational processes. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980. (ERIC Document Reproduction Service No. ED 199 269) (a)

Thompson, B. The instructional strategy decisions of teachers. Education, 1980, 101, 150-157. (b)

Thompson, B. Canonical correlation analysis: the most general data-analytic system In C. Goldsmith (Ed.), Proceedings of ISSUE '82. Chicago: ISSUE, Inc., 1982, pp. 462-480. [Order document #03991, National Auxillary Publication Service, P.O. Box 3513, Grand Central Station, New York, NY 10017]

Thompson, B., & Pitts, M.C. The use of factor adequacy coefficients. Journal of Experimental Education, 1981/82, 50, 101-104.

Thorndike, R.M. Canonical analysis and predictor selection. The Journal of Multivariate Behavioral Research, 1977, 12, 75-87.

Veldman, D.J. FORTRAN programming for the behavioral sciences. New York: Gardner, 1978.

van den Wollenberg, A.L. Redundancy analysis: an alternative for canonical correlation analysis. Psychometrika, 1977, 42, 207-219.

Table 1

Hypothetical Data Set

Case	X	Y	A	B
1	1(-0.72)	9(+1.36)	4(+0.08)	6(+0.33)
2	5(+1.90)	4(-0.15)	0(-1.02)	8(+0.88)
3	3(+0.59)	9(+1.36)	6(+0.64)	0(-1.33)
4	3(+0.59)	4(-0.15)	6(+0.64)	9(+1.16)
5	3(+0.59)	3(-0.45)	9(+1.46)	0(-1.33)
6	2(-0.07)	2(-0.76)	9(+1.46)	0(-1.33)
7	2(-0.07)	0(-1.36)	2(-0.47)	9(+1.16)
8	0(-1.38)	2(-0.76)	0(-1.02)	5(+0.06)
9	0(-1.38)	9(+1.36)	1(-0.74)	6(+0.33)
10	2(-0.07)	3(-0.45)	0(-1.02)	5(+0.06)

Note: Z score equivalents of the unstandardized data are presented in parentheses.

Table 2

Analyzed Matrix

.086	.011
.048	.027

Table 3

Calculation of Composites

Case	ZX	F1	ZY	F2	Criterion Composite	Predictor Composite	ZA	F3	ZB	F4
1	(-0.72)	(F1)+	(+1.36)	(F2) =	-0.97	+0.42 =	(+0.08)	(F3)+	(+0.33)	(F4)
2	(+1.90)	(F1)+	(-0.15)	(F2) =	+1.82	-0.49 =	(-1.02)	(F3)+	(+0.88)	(F4)
3	(+0.59)	(F1)+	(+1.36)	(F2) =	+0.26	-0.43 =	(+0.64)	(F3)+	(-1.33)	(F4)
4	(+0.59)	(F1)+	(-0.15)	(F2) =	+0.59	+1.92 =	(+0.64)	(F3)+	(+1.16)	(F4)
5	(+0.59)	(F1)+	(-0.45)	(F2) =	+0.65	+0.64 =	(+1.46)	(F3)+	(-1.33)	(F4)
6	(-0.07)	(F1)+	(-0.76)	(F2) =	+0.10	+0.64 =	(+1.46)	(F3)+	(-1.33)	(F4)
7	(-0.07)	(F1)+	(-1.36)	(F2) =	+0.23	+0.49 =	(-0.47)	(F3)+	(+1.16)	(F4)
8	(-1.38)	(F1)+	(-0.76)	(F2) =	-1.14	-1.27 =	(-1.02)	(F3)+	(+0.06)	(F4)
9	(-1.38)	(F1)+	(+1.36)	(F2) =	-1.59	-0.65 =	(-0.74)	(F3)+	(+0.33)	(F4)
10	(-0.07)	(F1)+	(-0.45)	(F2) =	+0.04	-1.27 =	(-1.02)	(F3)+	(+0.06)	(F4)

Note: The canonical function coefficients are respectively: "F1" = +0.94; "F2" = -0.22; "F3" = +1.30; "F4" = +0.94.

Table 4

Correlation Matrix

Variable		A	B	C	D	Q	R	S	T	U	V	W	X	Y	Z
INQSTRAT	(A)	—	00	00	00	15	10	-16	15	06	22	04	02	03	-12
INCISIVE	(B)	00	--	00	00	-24	24	05	07	13	06	17	04	06	-09
AFFECTIV	(C)	00	00	--	00	16	14	-10	07	-06	-07	02	19	-11	16
STRUCTUR	(D)	00	00	00	--	14	-05	17	-04	04	08	10	14	10	12
ESSENTIA	(Q)	-15	05	-10	17	--	00	00	00	00	00	00	00	00	00
HUMANISM	(R)	16	08	08	-02	00	--	00	00	00	00	00	00	00	00
PERENIAL	(S)	05	08	-13	10	00	00	--	00	00	00	00	00	00	00
PROGRESS	(T)	09	26	15	-04	00	00	00	--	00	00	00	00	00	00
RATIONAL	(U)	07	15	-07	04	00	00	00	00	--	00	00	00	00	00
EXISTENL	(V)	-10	-13	16	12	00	00	00	00	00	--	00	00	00	00
WARM	(W)	03	05	20	12	-01	05	-14	13	-04	-14	--	00	00	00
SCHOLARL	(X)	06	20	01	09	02	10	14	12	07	-14	00	--	00	00
RIGOROUS	(Y)	22	06	-09	10	08	05	15	-06	07	-01	00	00	--	00
IMPOTENT	(Z)	13	-26	18	15	01	04	-08	-07	-06	25	00	00	00	--

Note: Decimals omitted.

Table 5

Canonical Results

Variable	I	II	III	IV
INQSTRAT	-.38	-.66	.49	.43
INCISIVE	-.87	.03	-.42	-.25
AFFECTIV	.25	-.70	-.17	-.65
STRUCTUR	.18	-.28	-.75	.57
ESSENTIA	.07	.32	-.62	.24
HUMANISM	-.23	-.29	.20	-.10
PERENIAL	-.09	.03	-.21	.40
PROGRESS	-.48	-.32	-.05	-.41
RATIONAL	-.29	-.03	-.13	.19
EXISTENL	.35	-.13	-.50	-.25
WARM	.12	-.41	-.49	-.11
SCHOLARL	-.23	-.13	-.40	.02
RIGOROUS	-.30	-.29	.14	.52
IMPOTENT	.40	-.60	.17	.38
R_c^2	.210	.176	.113	.094

Table 6

Regression Solutions

Variable	I	II	III	IV	V	VI	VII	VIII	IX	X
INQSTRAT	-.15	.16	.05	.09	.07	-.10	.03	.06	.22	.13
INCISIVE	.05	.08	.08	.26	.15	-.13	.05	.20	.06	-.26
AFFECTIV	-.10	.08	-.13	.15	-.07	.16	.20	.01	-.09	.18
STRUCTUR	.17	-.02	.10	-.04	.04	.12	.12	.09	.10	.15
ESSENTIA	1.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
HUMANISM	.00	1.00	.00	.00	.00	.00	.00	.00	.00	.00
PERENIAL	.00	.00	1.00	.00	.00	.00	.00	.00	.00	.00
PROGRESS	.00	.00	.00	1.00	.00	.00	.00	.00	.00	.00
RATIONAL	.00	.00	.00	.00	1.00	.00	.00	.00	.00	.00
EXISTENL	.00	.00	.00	.00	.00	1.00	.00	.00	.00	.00
WARM	.00	.00	.00	.00	.00	.00	1.00	.00	.00	.00
SCHOLARL	.00	.00	.00	.00	.00	.00	.00	1.00	.00	.00
RIGOROUS	.00	.00	.00	.00	.00	.00	.00	.00	1.00	.00
IMPOTENT	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.00
R^2	.065	.039	.035	.098	.034	.069	.057	.054	.071	.138

Table 7

Non-symetric Technique Results

Variable	I	II	III	IV	V	VI	VII	VIII	IX	X
INQSTRAT	.15	.10	-.16	.15	.06	.22	.04	.02	.03	-.12
INCISIVE	-.24	.24	.05	.07	.13	.06	.17	.04	.06	-.09
AFFECTIV	.16	.14	-.10	.07	-.06	-.07	.02	.19	-.11	.16
STRUCTUR	.14	-.05	.17	-.04	.04	.08	.10	.14	.10	.12
ESSENTIA	.00	.00	1.00	.00	.00	-.04	-.01	.01	.01	.00
HUMANISM	-.02	.01	.00	1.01	.00	-.03	-.05	-.03	.00	.00
PERENIAL	.04	.00	.00	.00	.01	-.08	-.07	.07	1.03	-.01
PROGRESS	.04	1.02	.00	.00	.00	.03	-.06	-.07	.00	-.02
RATIONAL	.03	.00	.00	.00	1.01	-.04	-.04	.02	.01	-.01
EXISTENL	-.14	-.02	.00	.00	-.01	.00	.08	.07	-.01	1.04
WARM	-.01	-.07	.00	-.03	.02	-.01	.01	1.02	.07	.08
SCHOLARL	-.02	-.06	-.01	-.05	-.04	.01	1.03	.01	-.07	.08
RIGOROUS	.00	.03	-.04	-.03	-.04	1.02	.01	-.01	-.08	.01
IMPOTENT	1.03	.04	.00	-.02	.03	.00	-.02	-.01	.04	-.14
R^2	.123	.089	.066	.035	.028	.065	.040	.057	.027	.062

Figure 1

Scattergram of Canonical Composites